

Module 14: Missing Data

Stata Practical

Jonathan Bartlett & James Carpenter

London School of Hygiene & Tropical Medicine

www.missingdata.org.uk

Supported by ESRC grant RES 189-25-0103 and MRC grant G0900724

Pre-requisites

- Stata version 12 or later.

Online resources:

www.missingdata.org.uk

Contents

| | |
|--|-----------|
| Introduction to the Youth Cohort Study dataset | 1 |
| P14.1 The Model of Interest | 2 |
| P14.2 Investigating Missingness | 2 |
| P14.2.1 Investigating quantity and patterns of missingness | 2 |
| P14.2.2 Investigating the missingness mechanism | 4 |
| P14.3 Ad-hoc Methods | 7 |
| P14.4 Complete Records Analysis | 8 |
| P14.4.1 Complete records analysis results | 8 |
| P14.4.2 Interpretation of complete records analysis | 9 |
| P14.5 Multiple Imputation | 12 |
| P14.5.1 Imputation by Chained Equations in Stata | 12 |
| P14.5.2 Analysing the multiple imputations | 15 |
| P14.6 Inverse Probability Weighting | 18 |
| P14.6.1 Constructing the weights | 18 |

| | |
|---|-----------|
| P14.6.2 Inverse probability weighted complete records analysis..... | 20 |
| P14.7 Multilevel and Longitudinal Studies | 23 |
| P14.8 Summary and Conclusions | 24 |
| References..... | 25 |
| Acknowledgements..... | 25 |

Introduction to the Youth Cohort Study dataset

You will be analysing data from the Youth Cohort Study of England and Wales (YCS)¹. The YCS is a postal survey of young people. We will use data from the 1995 cohort, restricted to those young people who were at comprehensive schools (n=12,884) when the survey took place.

Our analyses will focus on variables recording GCSE attainment, parental socio-economic class, gender, and ethnicity. In particular our interest will focus on models for the young person's GCSE attainment score, with the other variables as covariates or explanatory variables. Such a model is of interest in order to investigate differences in GCSE attainment between ethnic and social economic groups, relative to gender differences (Connolly 2006). Table 14.1 describes the variables included in the dataset.

Table 14.1. Variables contained in Youth Cohort Study dataset

| Variable name | Description and coding |
|-----------------|---|
| t0score2 | GCSE score - truncated year 11 exam point score |
| gender | Gender (1 = boys, 0 = girls) |
| t0ethnic | Ethnicity 1 = white 4 = black 5 = Indian 6 = Pakistani 7 = Bangladeshi 9 = Other Asian 10 = other response |
| t0parsc4 | Parents' National Statistics Socio-economic classification 1 = managerial & professional 2 = intermediate 3 = working |

¹ We thank the depositors of the Economic and Social Data Service (ESDS) data collection SN 5765 'Youth Cohort Time Series for England, Wales and Scotland, 1984-2002', and the depositors of the constituent studies, for their permission to make these data available for teaching purposes. We also thank the ESDS (www.esds.ac.uk) for their assistance in obtaining these permissions and through whose website the data were made available to us.

The GCSE score is formed by assigning numerical scores to the grades obtained by a child at GCSE (A/A*=7 through to grade G=1), truncated at 12 grade A/A*s (giving a maximum score of 84).

The original YCS data also contains a weight variable, based on the sampling scheme used in the survey. Since our aim here is to illustrate the missing data concepts and methods we have introduced, we do not use the weights in this analysis. We emphasise that the analyses shown here are intended to be illustrative of the missing data concepts and methods we have introduced, and should not be interpreted as a substantive analysis of these data.

P14.1 The Model of Interest

Throughout the practical we shall assume that our model of interest is the linear regression of GCSE score on gender, ethnicity and parental SEC. Ordinarily we would fit this model in Stata using:


```
xi: regress t0score2 gender i.t0ethnic i.t0parsc4
```

We will keep this model of interest in mind when investigating missingness in the variables and when considering how to handle any missing values.

P14.2 Investigating Missingness

In this section we investigate missingness in the YCS data. Load “10.2.dta” into memory and open the do-file for this lesson:

From within the LEMMA Learning Environment

- Go to **Module 14: Missing Data**, and scroll down to  **Stata Datasets and Do-files**
- Click “14.2.dta” to open the dataset

P14.2.1 Investigating quantity and patterns of missingness

We begin by investigating how many missing values there are in the variables included in the dataset, using Stata’s `misstable` summarize command:

Module 14 (Practical): Missing Data in Stata

```
. misstable summarize t0score2 t0parsc4 t0ethnic gender
```

| Variable | Obs=. | Obs>. | Obs<. | Unique values | Min | Max |
|----------|-------|-------|--------|---------------|-----|-----|
| t0score2 | 129 | | 12,755 | 85 | 0 | 84 |
| t0parsc4 | 1,576 | | 11,308 | 3 | 1 | 3 |
| t0ethnic | 156 | | 12,728 | 7 | 1 | 10 |

We first note that the gender variable has not been included in the output - this is because the variable has no missing values. Next we see that the parental SEC has the most missing values (1,576), with GCSE score and ethnicity having fewer missing values.

Next we examine the patterns of missingness in these three variables. We use the `misstable patterns` command to tabulate which patterns of missingness occur and how frequently each pattern occurs:

```
. misstable patterns t0score2 t0parsc4 t0ethnic gender , freq
```

Missing-value patterns

(1 means complete)

| Frequency | Pattern |
|-----------|---------|
| | 1 2 3 |
| 11,188 | 1 1 1 |
| 1,422 | 1 1 0 |
| 104 | 1 0 0 |
| 77 | 0 1 1 |
| 41 | 0 1 0 |
| 41 | 1 0 1 |
| 9 | 0 0 0 |
| 2 | 0 0 1 |
| 12,884 | |

Variables are (1) t0score2 (2) t0ethnic (3) t0parsc4

The output from `misstable patterns` shows, for the specified variables, each pattern of missing data which occurs, ordered according to the frequency with which they occur. From the first row in the table, we see that there are 11,188 young people for whom all three variables (ethnicity, GCSE score, and parental SEC) are observed. The most common pattern which has some missing values is when GCSE score and ethnicity are observed but parental SEC is missing (n=1,422). The next most commonly occurring pattern is where GCSE score is observed but ethnicity and parental SEC are missing (n=104). We then see that all the other possible missingness patterns occur, but with smaller frequencies.

P14.2.2 Investigating the missingness mechanism

Since missingness occurs in three of the variables in the dataset, we can think of there being an underlying mechanism which determines missingness for each of the variables ethnicity, GCSE score, and parental SEC. Since the majority of missing values occur in the parental SEC variable, however, we shall focus on investigating missingness in this variable, since the analysis is likely most sensitive to assumptions concerning this. The other patterns of missing data we (implicitly) assume are either MCAR or possibly MAR given other observed values.

From the output from `misstable patterns`, we saw that when parental SEC is missing, ethnicity and GCSE score are mostly observed. We can therefore investigate how missingness in parental SEC is related both to these two variables and to the fully observed variable gender.

To investigate which variables are predictive of missingness in the parental SEC variable we first define a binary variable which indicates whether the parental occupation variable is observed (=1) or missing (=0):

```
gen r_t0parsc4=(t0parsc4!=.)
```

Next, we fit a logistic regression model for the variable `r_t0parsc4`, with `gender` as covariate (we could also have simply performed a chi-squared test):

```
. xi: logistic r_t0parsc4 gender
```

```
Logistic regression           Number of obs   =       12884
                              LR chi2(1)         =           1.21
                              Prob > chi2          =       0.2716
Log likelihood = -4786.1437    Pseudo R2       =       0.0001
```

| ----- | | | | | | |
|-------------|------------|-----------|-------|-------|----------------------|----------|
| r_t0parsc4 | Odds Ratio | Std. Err. | z | P> z | [95% Conf. Interval] | |
| -----+----- | | | | | | |
| gender | .9424493 | .0507889 | -1.10 | 0.271 | .8479816 | 1.047441 |
| ----- | | | | | | |

This document is only the first few pages of the full version.

To see the complete document please go to learning materials and register:
<http://www.cmm.bris.ac.uk/lemma>

The course is completely free. We ask for a few details about yourself for our research purposes only. We will not give any details to any other organisation unless it is with your express permission.